W350–W355 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh475

CysView: protein classification based on cysteine pairing patterns

Johann Lenffer[1,2], Paulo Lai[1,2], Wafaa El Mejaber[1,3], Asif M. Khan[1,4], Judice L. Y. Koh[1], Paul T. J. Tan[1,4], Seng H. Seah[1] and Vladimir Brusic[1,5,*]

[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, 119613 Singapore, [2]School of Molecular and Microbial Biosciences, Building G08, University of Sydney, NSW 2006, Australia, [3]IUP Génie Physiologique et Informatique, Bâtiment botanique, 40 av. Recteur Pineau-86022, Poitiers Cedex, France and [4]Department of Biochemistry and [5]Department of Microbiology, Faculty of Medicine, National University of Singapore, 8 Medical Drive, 117597 Singapore

Received February 15, 2004; Revised April 20, 2004; Accepted May 3, 2004

ABSTRACT

CysView is a web-based application tool that identifies and classifies proteins according to their disulfide connectivity patterns. It accepts a dataset of annotated protein sequences in various formats and returns a graphical representation of cysteine pairing patterns. CysView displays cysteine patterns for those records in the data with disulfide annotations. It allows the viewing of records grouped by connectivity patterns. CysView's utility as an analysis tool was demonstrated by the rapid and correct classification of scorpion toxin entries from GenPept on the basis of their disulfide pairing patterns. It has proved useful for rapid detection of irrelevant and partial records, or those with incomplete annotations. CysView can be used to support distant homology between proteins. CysView is publicly available at http://research.i2r.a-star.edu.sg/CysView/.

INTRODUCTION

Disulfide-containing proteins are diverse and include, among others, hydrolases, inhibitors, hormones and toxins. They constitute a large group of proteins with diverse functions, including cell-to-cell recognition, cell signalling and cell defence (1). Many of these proteins, such as toxins, exert their biological effects through specific interaction with various ion channels (2), membranes (3) and cellular receptors (4). These interactions are potent and specific, making disulfide-containing proteins attractive candidates for the development of drugs, new therapeutics and diagnostic agents (5,6).

Disulfide bonds (or bridges) are formed by oxidation of cysteine residues. These covalent bonds are the main cross-links present in proteins that play a key role in folding and stabilizing their three-dimensional (3D) scaffold (7–9). Disulfide bonds are also involved in the formation of certain structural motifs, such as the cystine knot in a number of proteins (10) and the CxxC motif in redox-active proteins (11). Cysteine bridges enhance the structural stability of proteins, and act as switches for protein function (12) and for regulation of enzymatic activity (9,13). Loss of disulfide bridges can lead to loss of both structural and functional characteristics of a protein. For example, replacement of cysteine pair residues with alanines in either of the two disulfide bridges in interleukin-8 (IL-8) results in loss of both structural integrity and function (14). Characterization of disulfide bonds includes prediction of the cysteine bonding state ('bonded' or 'free' cysteine) (15–17), prediction of disulfide connectivity from primary sequence for proteins lacking disulfide bond information (18,19), engineering of disulfide bonds (13), analysis of structure and sequence features (20–23) and classification of connectivity patterns (24). A database of disulfide bonds in proteins, DSDBASE, was recently created to provide information on native and modelled disulfide bonds in proteins (25).

Cysteine residue pairs form disulfide bridges in proteins, with each specific set of bridges within a protein termed a 'connectivity pattern'. Disulfide connectivity pattern formation in a protein is a directed (i.e. non-random) process (26). A large variety of these connectivity patterns can be found in disulfide-containing proteins (24,26). In proteins with low sequence similarity (i.e. sequence identity below 25%),

*To whom correspondence should be addressed at 21 Heng Mui Keng Terrace, 119613 Singapore. Tel: +65 6874 7920; Fax: +65 6774 8056; Email: vladimir@i2r.a-star.edu.sg

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

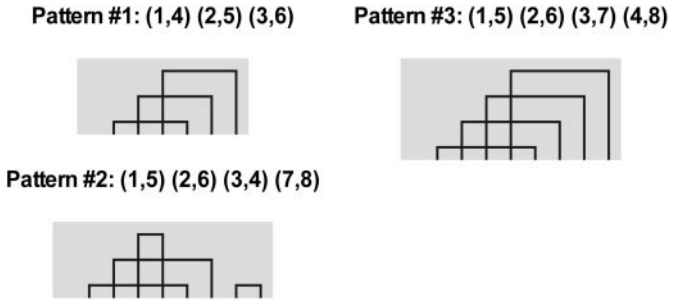


Figure 1. Examples of connectivity patterns in four scorpion toxins extracted by CysView. Scorpion toxins BMTX1 (Q9NII6) and LpII (P80670) share pattern 1, MTX (P80719) has pattern 2, and Pi4 (P58498) has pattern 3. MTX and Pi4 have a high sequence identity (81.6%) and the same number of cysteines, but different connectivity patterns; whereas scorpion toxins BMTX1 and LpII have a low sequence identity (21%) but share an identical connectivity pattern. All accession numbers refer to SwissProt entries.

identical connectivity patterns can indicate high structural homology (27). Likewise, high-similarity proteins do not necessarily share the same connectivity patterns (see examples in Figure 1). Proteins that share a disulfide bonding pattern usually belong to the same structurally derived family. Therefore, disulfide connectivity patterns provide a rapid and simple method for structural characterization of protein sequences and for examining structural properties, such as protein topologies (26), entropic effect of cross-linkage (24), structural superimposition of proteins by means of their disulfide bridge topology (28) and taxonomy of small disulfide-rich protein folds (29). In addition, identifying connectivity patterns might reveal interesting or even novel clustering within protein families, offering additional perspective on their overall structure–function relationship. For example, 3D structural superimposition of two apparently unrelated protein structures by means of their disulfide bridge topology facilitates determination of important structural or functional residues (1).

With the rapid expansion of sequence data stemming from both classical and genomic/proteomic approaches, identification and grouping of connectivity patterns manually from this large data pool is inefficient. Therefore, there is a need for a tool that automates the identification and classification process using connectivity patterns. Recently, van Vlijmen *et al*. (30) built a comprehensive database of disulfide bonding patterns and a search tool to identify proteins with similar disulfide patterns. However, this tool employs a similarity search for extraction of disulfide patterns stored in their database, but not of user-defined protein datasets retrieved from major public databases, such as GenPept (31), SwissProt (32) and PDB (Protein Data Bank) (33).

In this paper, we present CysView, a publicly available tool that identifies connectivity patterns present in an arbitrary group of proteins and that classifies proteins according to their disulfide connectivity patterns. It is a web-based tool that accepts a dataset of annotated protein sequences in several formats [full-text entries, a list of accession numbers or output from a protein–protein BLAST search, blastp, (34)] and returns a graphical representation of cysteine pairing patterns. It displays cysteine connectivity patterns for those records in the data that have disulfide annotations. It allows users to view either records grouped by connectivity patterns or records containing a given subpattern. It is also able to predict intra-disulfide connectivity patterns in proteins that lack disulfide bridging information. The availability of CysView is a useful addition to existing homology discovery methods. CysView is publicly accessible at http://research.i2r.a-star.edu.sg/CysView/.

SYSTEM DESCRIPTION

The CysView web interface allows two input methods: data can either be uploaded as a text file or entered directly into a textbox. The following data formats are supported (input files must be in the text file format, *.txt):

(i) full-text entries from the major public databases, GenPept and SwissProt. A list of full records from GenPept and SwissProt delimited by a line containing a compulsory '//';
(ii) full-text entries from databases built using BioWare software (http://research.i2r.a-star.edu.sg/Templar; J. L. Y. Koh, S. P. T. Krishnan, S. H. Seah, A. M. Khan, P. T. J. Tan, M. L. Lee and V. Brusic, manuscript in preparation);
(iii) a list of GenPept GIs (sequence identification numbers), SwissProt accessions or PDB IDs, one identifier (ID) per line;
(iv) output from a blastp search. Outputs from NCBI (http://www.ncbi.nlm.nih.gov/BLAST), ExPASy (http://www.expasy.org/tools/blast/) and DDBJ (http://www.ddbj.nig.ac.jp/E-mail/homology.html) (35) servers have been tested for CysView. Currently, only GenPept, SwissProt and DDBJ entries in the blastp file are allowed.

CysView requires full-text entries for the extraction of cysteine pairing information. For input consisting of a list of GIs, accession numbers or a BLAST result file, the corresponding entries from the SwissProt, GenPept, PDB and DDBJ databases are downloaded using a given list of IDs, which are extracted from the blastp results file if this is the input format (Figure 2). Duplicate entries within the input are handled by taking the first occurrence of a sequence and ignoring subsequent appearances of the repeat entries. However, we recommend that users submit clean datasets if possible. The process of defining connectivity patterns present in a local protein database is shown in Figure 2. A maximum of 500 PDB, 500 GenPept and 1000 SwissProt or DDBJ entries can be retrieved at any one time for accession number or BLAST results file inputs, due to batch-retrieval restrictions imposed by the originating databases.

Disulfide information is extracted from the full-text entries. This information for SwissProt entry P58752 is shown in Figure 3. To define the cysteine pairing description, a number is assigned to each disulfide pairing cysteine based on the order of appearance in the protein sequence, from N- to C-terminus. The entry P58752 has the cysteine pairing pattern (1,4) (2,5) (3,6). This means that the first cysteine (Cys18) pairs with the fourth cysteine (Cys41), the second (Cys27) pairs with the fifth (Cys46), and the third (Cys31) pairs with the sixth (Cys48).

The results page (Figure 4) starts with a summary of results. After the summary, connectivity patterns are listed in a table in descending order of number of occurrences.

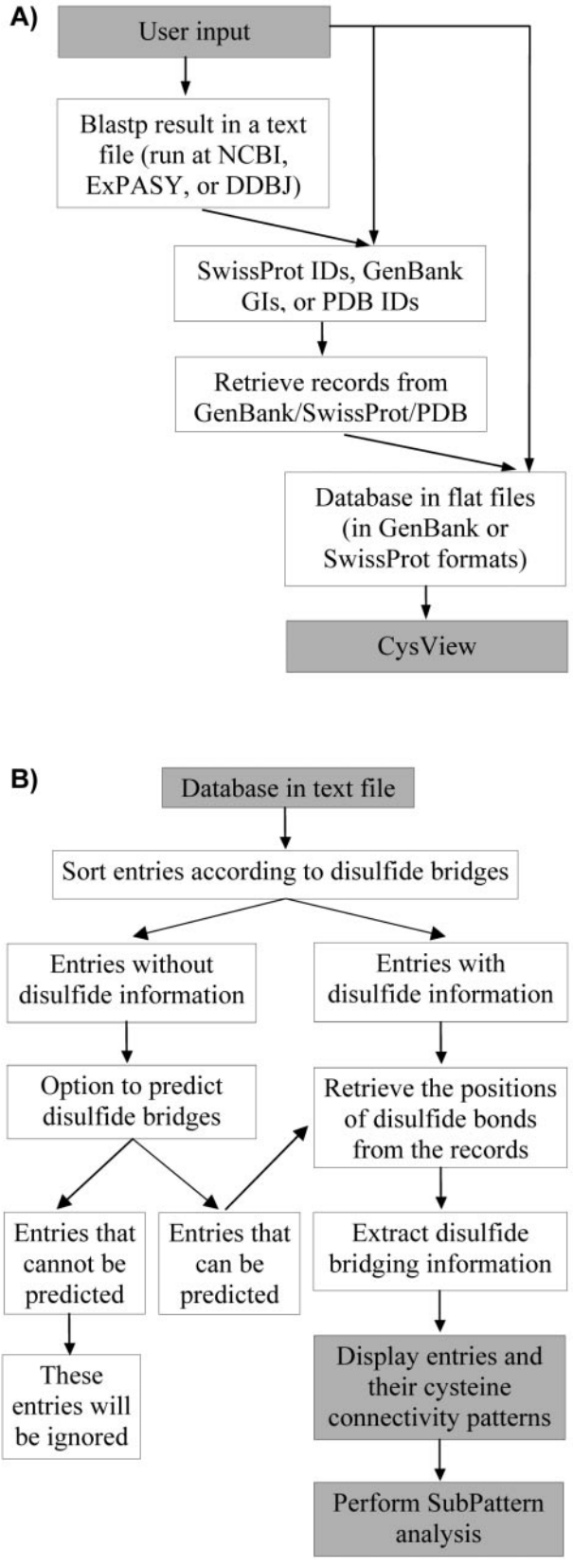


Figure 2. Data flow in CysView. (**A**) Progression of different input data formats through CysView. (**B**) The process of defining cysteine pairings present in a local protein database and prediction of cysteine pairings in sequences without that information.

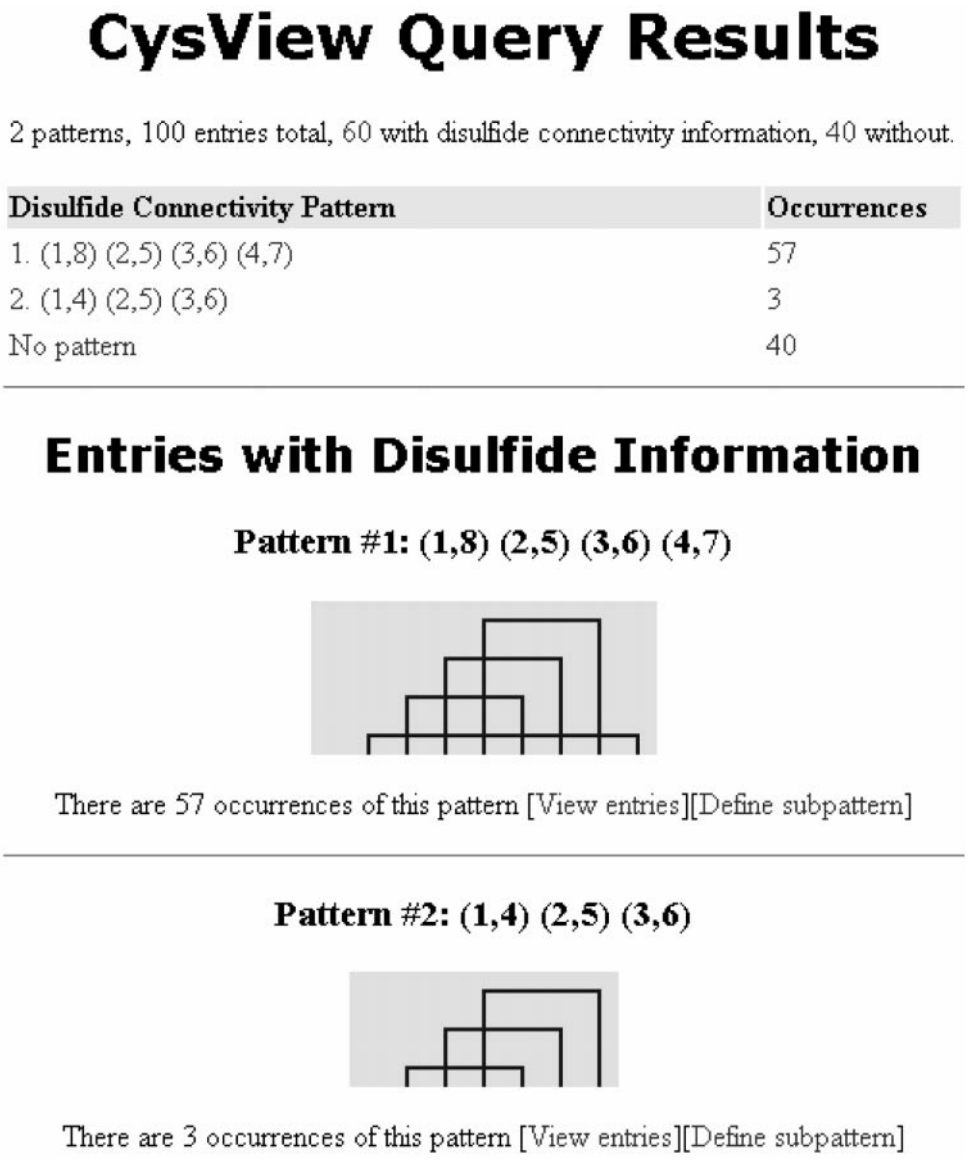
The entries without disulfide bond information are included at the bottom of this table. Each line in this table is hyperlinked to the selection of entries that share the same connectivity pattern. Below the connectivity pattern table is the visual representation of each pattern, listed in the same order as in the table. The visual representations are composed from a tile-set of six images; these represent a small part of the image such as corner line, blank space, vertical or horizontal line. The images are generated using Perl scripts that incorporate logic for converting textual description of features (Figure 3) into images (Figures 1 and 4). For data derived from blastp results, the percentage similarity to the query sequence is included alongside the entry's ID. Inter-monomer disulfide bonds in multi-subunit proteins fall outside the scope of this application.

CysView enables users to predict the connectivity pattern of a query sequence that has no disulfide bridge annotation. The connectivity pattern prediction module takes input in CysView-supported formats. The prediction of the connectivity pattern is performed by comparison of a query sequence to the set of 2350 SwissProt entries (release 43.1) that contain experimentally determined disulfide bridges. These entries contain 224 unique connectivity patterns (see supplementary material on the CysView homepage) whose sizes range from 1 bridge (882 sequences) to 36 bridges (1 sequence). The connectivity pattern is deduced from the most similar sequence (using BLAST search), and the number of cysteines differing at most by one between the query and target sequence. Entries that have their disulfide bridge predicted by the system are indicated by the term '(predicted)'. A 10-fold cross-validation result for the CysView disulfide connectivity prediction system indicates the positive predictive value of predictions to be PPV = 78.6%, which is comparable to the other disulfide connectivity prediction methods (18,19). The advantage of our method is that its predictive performance does not deteriorate with increasing number of disulfide bonds in the query sequence. However, our method can determine connectivity patterns only if they show similarity to any of the 2350 SwissProt entries that have experimentally determined disulfide bridges. A maximum of 320 query sequences can be processed by the prediction module at any one time.

Finally, CysView enables users to select a query subpattern from any existing connectivity patterns and then search to determine whether the query is contained within any other sequences available for analysis. The subpattern is defined by users, who select cysteines from an individual connectivity pattern. The output of the subpattern search is a list of connectivity patterns plus the list of relevant entries.

```
FT   DISULFID     18      41       BY SIMILARITY.
FT   DISULFID     27      46       BY SIMILARITY.
FT   DISULFID     31      48       BY SIMILARITY.
```

Figure 3. SwissProt features describing disulfide pairs from entry P58752. The numbers represent the protein positions of cysteine pairs that form disulfide bridges. In this entry, for example, the cysteine at position 18 pairs with the cysteine at position 41. Another two pairs are Cys27–Cys46 and Cys31–Cys48.



CysView Query Results

2 patterns, 100 entries total, 60 with disulfide connectivity information, 40 without.

Disulfide Connectivity Pattern	Occurrences
1. (1,8) (2,5) (3,6) (4,7)	57
2. (1,4) (2,5) (3,6)	3
No pattern	40

Entries with Disulfide Information

Pattern #1: (1,8) (2,5) (3,6) (4,7)

There are 57 occurrences of this pattern [View entries][Define subpattern]

Pattern #2: (1,4) (2,5) (3,6)

There are 3 occurrences of this pattern [View entries][Define subpattern]

Figure 4. The CysView results output generated from a blastp search of GenPept scorpion neurotoxin GI: 20140339 against the GenPept database. CysView proved useful in identifying records with potential errors in their disulfide bridge annotation. All three entries with pattern 2 were found to have incomplete annotations. This figure is the direct output from CysView.

CysView APPLICATIONS

CysView can be used for quick structural classification and preliminary analysis of a group of protein entries. To demonstrate the utility of CysView, we have applied it to a set of scorpion toxin entries and compared the results with groups defined in the SCORPION database (36).

Scorpion toxin entries were extracted from a major public database, GenPept, using the keyword search query 'Scorpion AND toxin', which returned a total of 538 entries (in February 2004). The retrieved dataset, without further processing, such as the removal of irrelevant records and duplicates, was then submitted to CysView for the analysis of cysteine connectivity patterns. CysView returned a total of 13 pattern groups from 315 entries which contained disulfide connectivity information (the output page is accessible from the CysView home page). To validate the patterns we inspected records within each pattern group. Of the groups, eight were valid (patterns 1, 2, 3, 4, 6, 7, 9 and 10), one was an unrelated false positive (pattern 11 was of a snake toxin, an alpha-cobratoxin), pattern 12 was a special case, as it was from an entry representing a two-subunit toxin, patterns 5 and 8 were partial entries, and pattern 13 represents a potential novel group (not yet defined in the SCORPION database). Partial entries refer to records with either incomplete sequence or incomplete disulfide annotation information. A BLAST search using members of pattern 5 indicates that they probably belong to group 2, whereas one member of pattern 8 actually belongs to group 7 and the others remain unknown. The novelty of group 13 (one member) can be validated once experimental data on the disulfide information become available. The results of CysView annotations were in excellent agreement with the

manually compiled connectivity pattern groups from the SCORPION database.

Any valid pattern group can then be the subject of further independent analysis to determine structural and functional properties. Sequence comparisons, phylogenetic studies, homology modelling and 3D structure analyses can help establish structure–function relationships within and between the groups.

The grouping of protein entries based on cysteine patterns is an important step for their structure–function classification and may support the prediction of biological activity directly from the classification schema (37). Furthermore, rich and well-organized data allow the use of sophisticated bioinformatics tools for improved structure–function analysis. In addition, we have shown how protein classification using the connectivity patterns facilitates the identification of unrelated, partial and novel entries retrieved from a database.

CysView has also proved useful in identifying records with potential errors in their disulfide bridge annotation. For example, when the result of a BLAST search using a GenPept scorpion neurotoxin (GI: 20140339) against GenPept was submitted to CysView, two patterns were returned (Figure 4), one with 57 members and the other with 3. The group of three consisted of two neurotoxins from *Centruroides exilicauda* (GI: 494705 and 4139618) and one from *Tityus serrulatus* (GI: 5821808). All three members of this group had their connectivity patterns validated manually. Multiple sequence alignment, literature review and crosschecking of each entry against its structure in the PDB database revealed that all three GenPept entries had an incomplete annotation. We have identified that these three entries were incorrectly classified as having the disulfide connectivity pattern (1,4) (2,5) (3,6) instead of (1,8) (2,5) (3,6) (4,7).

ACKNOWLEDGEMENTS

The authors would like to thank K. N. Srinivasan and two anonymous reviewers for their useful comments on the manuscript. J.L., P.L. and W.E.M. were international exchange students supported by the Institute for Infocomm Research, Singapore. A.M.K. and P.T.J.T. are graduate student scholars at the National University of Singapore.

REFERENCES

1. Mas,J.M., Aloy,P., Marti-Renom,M.A., Oliva,B., de Llorens,R., Aviles,F.X. and Querol,E. (2001) Classification of protein disulphide-bridge topologies. *J. Comput. Aided Mol. Des.*, **15**, 477–487.
2. Possani,L.D., Becerril,B., Delepierre,M. and Tytgat,J. (1999) Scorpion toxins specific for Na^+-channels. *Eur. J. Biochem.*, **264**, 287–300.
3. Hains,P.G., Ramsland,P.A. and Broady,K.W. (1999) Modeling of acanthoxin A1, a PLA2 enzyme from the venom of the common death adder (*Acanthophis antarcticus*). *Proteins*, **35**, 80–88.
4. Valentin,E. and Lambeau,G. (2000) What can venom phospholipases A(2) tell us about the functional diversity of mammalian secreted phospholipases A(2)? *Biochimie*, **82**, 815–831.
5. Harvey,A.L. (2002) Toxins 'R' Us: more pharmacological tools from nature's superstore. *Trends Pharmacol. Sci.*, **23**, 201–203.
6. Harvey,A.L., Bradley,K.N., Cochran,S.A., Rowan,E.G., Pratt,J.A., Quillfeldt,J.A. and Jerusalinsky,D.A. (1998) What can toxins tell us for drug discovery? *Toxicon*, **36**, 1635–1640.
7. Narayan,M., Welker,E., Wedemeyer,W.J. and Scheraga,H.A. (2000) Oxidative folding of proteins. *Acc. Chem. Res.*, **33**, 805–812.
8. Wedemeyer,W.J., Welker,E., Narayan,M. and Scheraga,H.A. (2000) Disulfide bonds and protein folding. *Biochemistry*, **39**, 4207–4216.
9. Kadokura,H., Katzen,F. and Beckwith,J. (2003) Protein disulfide bond formation in prokaryotes. *Annu. Rev. Biochem.*, **72**, 111–135.
10. Craik,D.J., Daly,N.L. and Waine,C. (2001) The cystine knot motif in toxins and implications for drug design. *Toxicon*, **39**, 43–60.
11. Kobayashi,T. and Ito,K. (1999) Respiratory chain strongly oxidizes the CXXC motif of DsbB in the *Escherichia coli* disulfide bond formation pathway. *EMBO J.*, **18**, 1192–1198.
12. Hogg,P.J. (2003) Disulfide bonds as switches for protein function. *Trends Biochem. Sci.*, **28**, 210–214.
13. Matsumura,M., Signor,G. and Matthews,B.W. (1989) Substantial increase of protein stability by multiple disulphide bonds. *Nature*, **342**, 291–293.
14. Rajarathnam,K., Sykes,B.D., Dewald,B., Baggiolini,M. and Clark-Lewis,I. (1999) Disulfide bridges in interleukin-8 probed using non-natural disulfide analogues: dissociation of roles in structure from function. *Biochemistry*, **38**, 7653–7658.
15. Martelli,P.L., Fariselli,P., Malaguti,L. and Casadio,R. (2002) Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy. *Protein Sci.*, **11**, 2735–2739.
16. Mucchielli-Giorgi,M.H., Hazout,S. and Tuffery,P. (2002) Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins*, **46**, 243–249.
17. Sharma,D. and Rajarathnam,K. (2000) 13C NMR chemical shifts can predict disulfide bond formation. *J. Biomol. NMR*, **18**, 165–171.
18. Vullo,A. and Frasconi,P. (2004) Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics*, **20**, 653–659.
19. Fariselli,P. and Casadio,R. (2001) Prediction of disulfide connectivity in proteins. *Bioinformatics*, **17**, 957–964.
20. Thornton,J.M. (1981) Disulphide bridges in globular proteins. *J. Mol. Biol.*, **151**, 261–287.
21. Petersen,M.T., Jonson,P.H. and Petersen,S.B. (1999) Amino acid neighbours and detailed conformational analysis of cysteines in proteins. *Protein Eng.*, **12**, 535–548.
22. Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.
23. Srinivasan,N., Sowdhamini,R., Ramakrishnan,C. and Balaram,P. (1990) Conformations of disulfide bridges in proteins. *Int. J. Pept. Protein. Res.*, **36**, 147–155.
24. Harrison,P.M. and Sternberg,M.J. (1994) Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage. *J. Mol. Biol.*, **244**, 448–463.
25. Vinayagam,A., Pugalenthi,G., Rajesh,R. and Sowdhamini,R. (2004) DSDBASE: a consortium of native and modelled disulphide bonds in proteins. *Nucleic Acids Res.*, **32**, D200–D202.
26. Benham,C.J. and Jafri,M.S. (1993) Disulfide bonding patterns and protein topologies. *Protein Sci.*, **2**, 41–54.
27. Chuang,C.C., Chen,C.Y., Yang,J.M., Lyu,P.C. and Hwang,J.K. (2003) Relationship between protein structures and disulfide-bonding patterns. *Proteins*, **53**, 1–5.
28. Mas,J.M., Aloy,P., Marti-Renom,M.A., Oliva,B., Blanco-Aparicio,C., Molina,M.A., de Llorens,R., Querol,E. and Aviles,F.X. (1998) Protein similarities beyond disulphide bridge topology. *J. Mol. Biol.*, **284**, 541–548.
29. Harrison,P.M. and Sternberg,M.J. (1996) The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J. Mol. Biol.*, **264**, 603–623.
30. van Vlijmen,H.W., Gupta,A., Narasimhan,L.S. and Singh,J. (2004) A novel database of disulfide patterns and its application to the discovery of distantly related homologs. *J. Mol. Biol.*, **335**, 1083–1092.
31. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
32. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., *et al.*

(2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
33. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
34. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
35. Tateno,Y., Miyazaki,S., Ota,M., Sugawara,H. and Gojobori,T. (2000) DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res*., **28**, 24–26.
36. Srinivasan,K.N., Gopalakrishnakone,P., Tan,P.T., Chew,K.C., Cheng,B., Kini,R.M., Koh,J.L., Seah,S.H. and Brusic,V. (2002) SCORPION, a molecular database of scorpion toxins. *Toxicon*, **40**, 23–31.
37. Tan,P.T., Khan,A.M. and Brusic,V. (2003) Bioinformatics for venom and toxin sciences. *Brief. Bioinform*., **4**, 53–62.